Faculty of Science, Technology, Engineering and Mathematics
M140 Introducing statistics

M140

TMA 03 2019B

Covers Units 6, 7, 8 and 9 **Cut-off date 17 July 2019**

Please read the Assessment Guide on the module website before beginning work on this TMA. You can submit your TMA either by post or electronically using the University's online TMA/EMA service.

This TMA is marked out of 100. Your overall score for this TMA will be the sum of your marks for each question.

The marks allocated to each part of each question are indicated in brackets in the margin.

Guidance about how to answer TMA questions is given in Subsection 7.2 of Unit 1.

Note that the Minitab files that you require for this assignment should be downloaded from the 'Assessment' area on the module website.

Copyright © 2018 The Open University
WEB 06683 3
12.1

You should be able to answer Questions 1 and 2 after you have studied Unit 6. You will need to use Minitab to answer Question 2.

You should be able to answer Questions 3 and 4 after you have studied Unit 7.

You should be able to answer Questions 5 and 6 after you have studied Unit 8. You will need to use Minitab to answer Question 6.

You should be able to answer Questions 7 and 8 after you have studied Unit 9.

Question 1 (Unit 6) – 18 marks

(a) Table 1 shows the distribution of women (in thousands) in England and Wales according to their marital status in mid-1957.

Table 1

	Marital status			
Age (years)	Single	Married	Widowed or divorced	Total
15–19	1306	83	0	1389
20–24	619	765	3	1387
25–29	263	1194	9	1466
30–34	173	1372	28	1573
35–39	171	1393	51	1615
40–44	159	1372	81	1612
45–49	208	1350	108	1666
50 or over	1116	4100	2329	7545
Total	4015	11629	2609	18253

Suppose that a woman is selected at random from this population.

(i) Calculate the probability that the selected woman is aged 35–39 years. [*2*]

(ii) Calculate the probability that the selected woman is married and aged 40–44 years. [*2*]

(iii) Calculate the probability that the selected woman is not single. You should perform this calculation in two different ways, one of which uses the probability rule for complementary events, and one that does not. Show your working. [*4*]

(b) Three friends, Helga, Amelie and Sun, regularly meet at a café. Based on previous experience, the probability that Helga buys a slice of cake is 0.45, the probability that Amelie buys a slice of cake is 0.38, and the probability that Sun buys a slice of cake is 0.70. What is the probability that all three buy a slice of cake the next time they visit the café? State any important assumption that you make in order to calculate this probability. [*3*]

(c) A random sample of size 9 is selected from a certain population.

(i) Calculate, by hand, the probability that exactly eight of the nine selected values lie above the population median. Show your working. [3]

(ii) Calculate, by hand, the probability that at least eight of the nine selected values lie above the population median. Again, show your working. [4]

Question 2 (Unit 6) – 7 marks

The Minitab file that you require for this question should be downloaded from the 'Assessment' area on the module website.

This question concerns a study to investigate whether the order of birth in pairs of twins determines which twin tends to be the more aggressive of the two. Each twin is scored for aggressiveness using a range of psychological tests, a higher score indicating greater aggressiveness. The Minitab worksheet **aggression.mtw** contains two columns.

- In the first column, headed `First born`, are the aggressiveness scores recorded for the first born in each of 12 pairs of twins.
- In the second column, headed `Second born`, are the aggressiveness scores recorded for the second born in each of *the same* 12 pairs of twins.

Run Minitab and open the worksheet.

(a) Use Minitab's **Calculator** facility to work out the differences between the values recorded for the first born and second born twins, and put the resulting differences in a new column called `difference`; that is, set `difference` = `First born` − `Second born`. List the values of `difference` as your answer to this part of the question. [1]

(b) In part (c) you will use Minitab to perform a sign test to address the question of whether or not the order of birth of the twins affects their levels of aggressiveness. Specify an appropriate hypothesis that the sign test can be used to test. [1]

(c) Use Minitab to perform an appropriate sign test on the differences. Include a copy of the relevant Minitab output in your answer. [2]

(d) Give the p-value from the test performed in part (c), and say what may be concluded from the test in terms of evidence about the hypothesis. Relate your conclusion back to consideration of the order of birth of twins and their aggressiveness levels. [3]

Question 3 (Unit 7) – 15 marks

(a) For the normal distribution shown in Figure 1, find approximate values for its mean and standard deviation. Explain how you obtained your answers. [3]

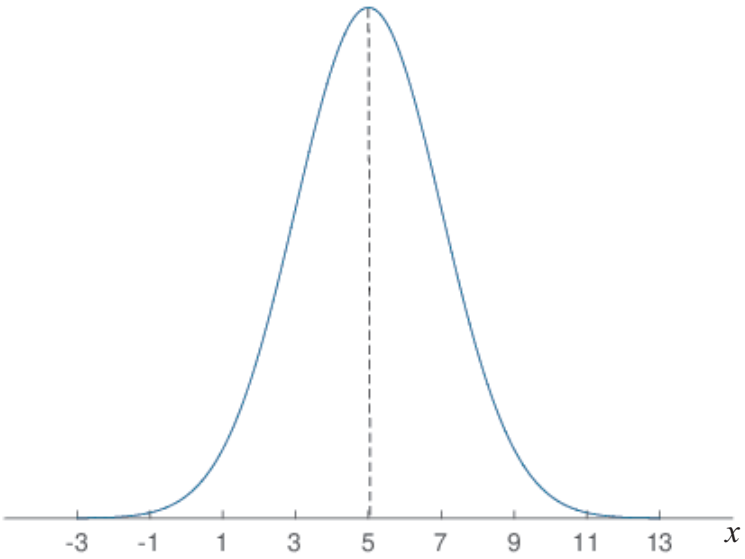

Figure 1

(b) (i) The normal distribution of a variable x has mean $\mu = -1$ and standard deviation $\sigma = 0.5$. Sketch this distribution by hand. [3]

(ii) Write down the formula for z that converts each value of the variable x in part (b)(i) to the number of standard deviations from its mean. Your formula should be written down in a form that does not involve any fraction or decimal. [2]

(iii) Calculate the value of z corresponding to $x = -1.8$. Interpret the value of z in terms of the number of standard deviations x is above or below its mean. [3]

(c) Emails sent internally within a large organisation may or may not have extra files associated with them as 'attachments'. Consider only emails that have a single attachment. Suppose that the population distribution of the sizes of such attachments has a mean of 0.95 megabytes (MB) and a standard deviation of 0.4 MB.

(i) Find the standard deviation of the sampling distribution of the mean for samples of 30 such attachments. [2]

(ii) Hence give the approximate distribution of the sample mean for samples of 30 such attachments. [2]

Question 4 (Unit 7) – 10 marks

A high street store was interested to discover if the spending of customers using credit cards is different from the spending of customers using cash. A random sample of 52 customers using credit cards for a single transaction had a mean spend of £36.43 with standard deviation £12.08. A random sample of 38 customers using cash for a single transaction had a mean spend of £31.84 with standard deviation £11.27.

A hypothesis test is to be performed to investigate whether the mean spend by customers using credit cards was equal to the mean spend by customers using cash.

(a) Name the hypothesis test that it is appropriate to use in this situation. [*1*]

(b) Using appropriate notation, which you should define, specify the null and alternative hypotheses associated with the test. [*2*]

(c) Calculate the value of the estimated standard error of the difference between the sample means. [*2*]

(d) Calculate the value of the test statistic. [*2*]

(e) Complete the hypothesis test, carefully detailing the conclusions of the test. [*3*]

Question 5 (Unit 8) – 15 marks

A stall sells cups of coffee, which may be caffeinated or decaffeinated, and may contain milk or be milk-free. It turns out that:

- 60% of the cups of coffee sold at the stall contain milk;
- of those cups of coffee sold at the stall containing milk, 24% are also decaffeinated.

(a) Let A denote the event that a cup of coffee sold at the stall contains milk, and let B denote the event that a cup of coffee sold at the stall is decaffeinated. Write the information given in the two bullet points above using probabilities in symbolic form. [*2*]

(b) Calculate the probability that a randomly selected coffee sold at the stall both contains milk and is decaffeinated. [*3*]

(c) Additional information is now given that 35% of coffees sold at the stall are decaffeinated. Calculate the probability that a randomly selected coffee sold at the stall that is decaffeinated also contains milk. [*4*]

(d) What percentage (to the nearest whole per cent) of coffees sold at the stall either contain milk, or are decaffeinated, or both? [*4*]

(e) For coffees sold at this stall, are the events of containing milk and being decaffeinated independent? Give a reason for your answer. [*2*]

Question 6 (Unit 8) – 15 marks

The Minitab file that you require for this question should be downloaded from the 'Assessment' area on the module website.

A study was made in five US hospitals of the effectiveness of a surgical procedure designed to improve the functioning of certain joints impaired by disease. Effectiveness was carefully defined and categorised into three levels, 'no improvement', 'partial improvement' and 'complete improvement'; the five hospitals were called A, B, C, D and E. No patient underwent more than one such surgical procedure.

The Minitab worksheet **hospital-surgery.mtw** contains the data for this study. In addition to a first column containing labels, the worksheet contains five columns of numbers, one associated with each hospital. It also contains three rows, one associated with each level of improvement. The data in the worksheet are the numbers of patients in each hospital and level of improvement category. Run Minitab and open the worksheet.

The data in **hospital-surgery.mtw** are to be analysed using a χ^2 test for contingency tables.

(a) Explain why the table in **hospital-surgery.mtw** is indeed a contingency table. [*4*]

(b) Specify the null and alternative hypotheses associated with the χ^2 test. [*2*]

(c) Perform the χ^2 test using Minitab. Include a copy of the relevant Minitab output in your answer. Identify the expected value for the patients in hospital C for whom the procedure resulted in partial improvement. [*2*]

(d) The smallest observed value in the table, that for patients in hospital B for whom the procedure resulted in no improvement, is 5. The χ^2 test that you carried out in part (c) remains valid. Explain why the first sentence in this part of the question is irrelevant to the second, and give the real reason for the validity of the χ^2 test in this case. [*2*]

(e) Minitab gives the degrees of freedom associated with this χ^2 test as 8. Show how this value arises. [*2*]

(f) Interpret the results of the χ^2 test, giving your conclusions about the relationship between hospital and effectiveness of surgical procedure. [*3*]

Question 7 (Unit 9) – 10 marks

Figure 2 shows a scatterplot of x and y values for each of 100 data points in a sample.

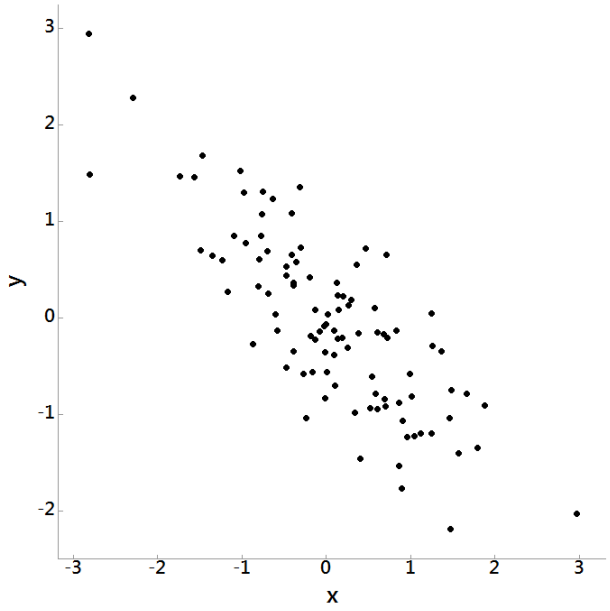

Figure 2

The correlation coefficient associated with Figure 2 is claimed to be -0.9.

(a) Is the value claimed for the correlation coefficient plausible in terms of its sign? Justify your answer. [2]

(b) Is the value claimed for the correlation coefficient plausible in terms of its closeness to -1? Justify your answer. [2]

In parts (c), (d) and (e), further statements are made about the relationship between x and y in Figure 2. In each case, state whether or not the statement is correct, giving a reason for your answer.

(c) If each value of the y variable is doubled, then the value of the correlation coefficient will be halved. [2]

(d) If the values of x and y are swapped over, then this will change the sign of the correlation coefficient. [2]

(e) Figure 2 shows that the variable x causes the variable y to take the values that it does. [2]

Question 8 (Unit 9) – 10 marks

The weights of 40 bags of sugar from a certain producer, labelled as containing 1 kg each, were measured. The sample mean weight of the bags of sugar was 1.0042 kg, and the sample standard deviation was 0.042 kg.

(a) Calculate the value of the estimated standard error of the sample mean. [*2*]

(b) Calculate a 99% confidence interval for the population mean weight of bags of sugar labelled 1 kg from this producer. [*3*]

(c) Interpret the confidence interval from part (b) in terms of all possible random samples of bags of sugar labelled 1 kg from this producer. [*2*]

(d) On the basis of the confidence interval from part (b), what would have been the outcome of a z-test of the null hypothesis that the population mean of bags of sugar labelled 1 kg from this producer is in fact 1.03 kg? Interpret the result of the test. (Note that you should use the confidence interval from part (b) to answer this question, and you should *not* perform a z-test.) [*3*]